Panel Data Analysis: Home Assignment 1

Comparative Social Research

Kirill Gasanovich Guliev

25 Feb 2022

knitr::opts\_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)  
  
path <- "C:/Users/Кирилл/Desktop/CSR2023/panel\_data/panel-data-analysis/"

library(haven)  
library(tidyverse)  
library(margins)  
library(psych)

# Problem 1. Load and subset data. Wave = 3

wvs\_data3 <- read\_dta(paste0(path, "data/soc\_hw1.dta")) %>%   
 filter(wave == 3) %>%   
 mutate(index = as.character(index),  
 wave = as.character(wave),  
 sex = as.factor(sex),  
 emp = as.factor(emp))

# Problem 2. Descriptive statistics

## Report descriptive statistics of all the variables

summary(wvs\_data3)

## index wave country satis   
## Length:76187 Length:76187 Length:76187 Min. : 1.000   
## Class :character Class :character Class :character 1st Qu.: 5.000   
## Mode :character Mode :character Mode :character Median : 7.000   
## Mean : 6.338   
## 3rd Qu.: 8.000   
## Max. :10.000   
## NA's :2734   
## chldr fsatis sex age emp   
## Min. :0.000 Min. : 1.000 0 :39542 Min. :15.00 0 :31830   
## 1st Qu.:0.000 1st Qu.: 3.000 1 :36588 1st Qu.:28.00 1 :40458   
## Median :2.000 Median : 5.000 NA's: 57 Median :39.00 NA's: 3899   
## Mean :1.843 Mean : 5.234 Mean :40.99   
## 3rd Qu.:3.000 3rd Qu.: 7.000 3rd Qu.:52.00   
## Max. :8.000 Max. :10.000 Max. :95.00   
## NA's :3754 NA's :4887 NA's :177

We see that in our filtered dataset we have 76’187 observations. We have 39’542 females and 36’588 males in our sample; there also 57 missing values concerning the question on respondent’s sex. The minimum age of the respondent is 15, the maximum is 95. The mean age in the sample is 40.99; 177 respondent’s have a missing value in their age. 40’458 respondents from the sample are employed, 31’830 are not employed. Variable on employment has the greatest number of missing values across the wave - 3’899.

The mean value of life satisfaction (satis variable) is 6.338. The median is 7 - which means that 50% of the sample estimate their life satisfaction above this point, and 50% - below. The financial satisfaction seems to be the most balanced variable - a half of sample estimated it above 5/10, and another half - below.

The maximum number of children in a family is 8, minimum is 0; 3’754 values are missing. 50% of the sample have more than 2 children.

table(wvs\_data3$country, useNA = "ifany")

##   
## Albania Argentina Armenia Australia   
## 22233 999 1070 1970 1974   
## Azerbaijan Belarus Brazil Chile China   
## 1963 1999 1143 999 1472   
## Colombia Estonia Georgia Germany India   
## 6004 981 2008 2002 2037   
## Japan Mexico New Zealand Nigeria Pakistan   
## 1031 1505 1177 1969 733   
## Peru Philippines Poland Romania Russia   
## 1120 1198 1153 1190 2020   
## South Africa South Korea Spain Sweden Taiwan   
## 2935 1248 1209 990 767   
## Turkey Ukraine United States Uruguay   
## 1865 2699 1525 999

There are 22’233 unlabeled observations in our Wave 3 of the WVS dataset.

# Distribution of number of children across countries  
table(wvs\_data3$country, wvs\_data3$chldr, useNA = "ifany")

##   
## 0 1 2 3 4 5 6 7 8 <NA>  
## 6041 3815 7160 2968 1086 467 218 140 211 127  
## Albania 216 88 262 236 137 31 10 3 2 14  
## Argentina 311 151 245 185 92 36 22 12 14 2  
## Armenia 653 193 515 408 121 54 20 2 2 2  
## Australia 679 201 497 342 153 53 20 8 15 6  
## Azerbaijan 663 244 464 319 157 61 27 19 9 0  
## Belarus 364 500 808 213 63 32 7 4 2 6  
## Brazil 370 176 241 170 76 46 17 18 29 0  
## Chile 250 163 227 172 94 45 18 10 15 5  
## China 262 431 417 180 94 45 18 3 7 15  
## Colombia 873 518 632 405 241 130 73 42 82 3008  
## Estonia 200 307 343 93 31 3 2 1 1 0  
## Georgia 618 381 657 268 63 14 6 1 0 0  
## Germany 591 467 649 200 55 20 8 3 3 6  
## India 439 249 374 401 290 140 43 23 11 67  
## Japan 240 122 420 193 36 10 2 0 1 7  
## Mexico 433 203 261 168 147 98 59 30 48 58  
## New Zealand 255 133 301 245 134 49 21 8 8 23  
## Nigeria 815 133 225 221 216 162 113 34 50 0  
## Pakistan 185 55 61 81 84 74 49 42 58 44  
## Peru 347 179 192 164 80 53 38 21 26 20  
## Philippines 374 126 171 169 153 74 53 29 49 0  
## Poland 253 188 344 192 90 41 21 11 11 2  
## Romania 312 334 357 111 45 12 5 0 1 13  
## Russia 295 642 821 180 48 23 2 3 1 5  
## South Africa 686 441 646 503 297 155 76 47 55 29  
## South Korea 352 165 365 177 89 33 15 11 3 38  
## Spain 391 173 306 173 87 36 15 14 9 5  
## Sweden 281 140 350 157 40 8 3 1 0 10  
## Taiwan 163 59 181 206 94 34 21 7 2 0  
## Turkey 317 252 405 273 161 87 49 45 61 215  
## Ukraine 411 736 1202 251 54 22 4 3 1 15  
## United States 340 202 405 260 168 75 25 20 24 6  
## Uruguay 241 169 282 153 66 40 19 7 16 6

Colombia has the most quantity of missing values concerning the variable on the number of children in the family - 3’008.

# Distribution of life satisfaction estimates across countries  
table(wvs\_data3$country, wvs\_data3$satis, useNA = "ifany")

##   
## 1 2 3 4 5 6 7 8 9 10 <NA>  
## 1178 747 1473 1547 3394 2338 2840 3735 2112 2681 188  
## Albania 31 70 140 240 187 124 114 79 8 0 6  
## Argentina 36 24 28 43 169 101 182 216 80 188 3  
## Armenia 283 208 325 251 335 182 160 109 57 49 11  
## Australia 13 15 50 50 169 122 316 595 333 302 9  
## Azerbaijan 126 70 187 257 495 181 200 229 96 83 39  
## Belarus 251 178 322 255 512 139 143 108 38 33 20  
## Brazil 56 34 41 43 150 100 110 171 98 336 4  
## Chile 19 9 31 57 147 162 140 187 93 151 3  
## China 50 45 62 69 174 196 210 254 165 242 5  
## Colombia 104 49 69 105 276 303 520 1250 1055 2252 21  
## Estonia 89 45 126 110 244 110 103 89 39 25 1  
## Georgia 357 122 209 220 394 200 157 165 80 93 11  
## Germany 26 18 73 108 237 248 358 562 223 146 3  
## India 133 57 107 110 226 216 268 297 217 291 115  
## Japan 16 16 39 44 145 201 177 242 74 52 25  
## Mexico 30 15 37 55 110 121 175 313 277 352 20  
## New Zealand 15 8 33 27 95 76 166 286 197 245 29  
## Nigeria 105 69 110 148 190 223 261 320 236 300 7  
## Pakistan 0 0 0 0 0 0 0 0 0 0 733  
## Peru 49 28 51 65 217 165 155 131 76 164 19  
## Philippines 29 22 41 63 250 143 124 182 121 222 1  
## Poland 49 34 49 71 235 137 155 165 87 160 11  
## Romania 135 115 126 153 203 132 118 108 63 32 5  
## Russia 325 171 293 231 417 131 134 156 56 83 23  
## South Africa 253 120 193 256 453 264 300 448 224 416 8  
## South Korea 0 0 0 0 0 0 0 0 0 0 1248  
## Spain 15 11 46 64 219 191 349 27 182 97 8  
## Sweden 6 5 17 31 68 53 160 282 193 173 2  
## Taiwan 14 4 33 34 130 170 126 134 52 68 2  
## Turkey 118 39 93 108 495 177 228 233 74 293 7  
## Ukraine 499 257 419 336 524 183 124 117 47 58 135  
## United States 18 21 27 42 123 105 238 371 271 301 8  
## Uruguay 24 10 36 51 130 102 153 190 97 202 4

Respondents from Pakistan and South Korea do not have values on life satisfaction.

# Distribution of financial satisfaction across countries  
table(wvs\_data3$country, wvs\_data3$fsatis, useNA = "ifany")

##   
## 1 2 3 4 5 6 7 8 9 10 <NA>  
## 2762 1342 2199 2023 3455 2311 2113 2212 951 1614 1251  
## Albania 32 85 217 181 127 147 117 65 11 2 15  
## Argentina 169 49 68 112 217 142 145 88 19 51 10  
## Armenia 486 262 281 249 254 178 129 81 18 19 13  
## Australia 83 50 101 141 317 198 374 321 128 250 11  
## Azerbaijan 231 116 270 287 540 130 119 136 56 45 33  
## Belarus 549 259 378 226 335 88 72 49 15 13 15  
## Brazil 168 52 80 84 214 131 108 97 34 169 6  
## Chile 59 20 57 95 203 159 155 130 49 71 1  
## China 78 60 88 96 246 228 221 195 93 159 8  
## Colombia 33 14 49 70 162 241 393 664 531 836 3011  
## Estonia 200 73 193 125 164 80 71 41 10 23 1  
## Georgia 784 194 302 187 247 105 81 50 18 32 8  
## Germany 87 41 119 158 361 281 317 352 108 171 7  
## India 93 60 145 215 293 188 236 277 155 219 156  
## Japan 27 14 62 60 131 222 157 214 55 53 36  
## Mexico 51 21 52 70 120 141 195 313 196 317 29  
## New Zealand 60 24 61 80 157 142 189 199 73 157 35  
## Nigeria 207 107 170 166 242 231 224 245 160 192 25  
## Pakistan 119 114 70 53 87 70 68 79 35 38 0  
## Peru 140 40 94 97 232 171 123 98 25 73 27  
## Philippines 77 34 67 79 288 152 151 126 79 144 1  
## Poland 210 97 139 133 227 104 112 55 17 52 7  
## Romania 278 127 169 134 192 106 90 56 8 22 8  
## Russia 658 215 303 214 337 105 72 46 14 46 10  
## South Africa 663 188 274 264 390 242 238 275 93 294 14  
## South Korea 47 30 87 108 340 202 193 146 38 47 10  
## Spain 47 38 89 105 290 219 276 15 62 58 10  
## Sweden 49 30 70 80 134 103 175 182 55 107 5  
## Taiwan 23 19 37 37 166 189 98 113 27 54 4  
## Turkey 131 65 153 147 632 258 205 151 24 92 7  
## Ukraine 887 343 484 288 303 115 74 51 21 32 101  
## United States 83 34 82 97 197 159 255 257 121 234 6  
## Uruguay 37 21 54 66 146 111 141 159 58 200 6

Colombia has the most quantity of missing values - 3’011.

## Explain why these preliminary statistics are especially important before running regression models with interaction terms

It is important to do exploratory analysis of the data before running linear models with interactions to see, what variables are balanced (have approximately the same proportions and normally distributed), where to expect outliers (which can influence the effect of independent variables on a dependent). Moreover, we need to bind to linear models assumptions one of which is normality of distribution of the residuals. It is more likely that residuals will be normally distributed if we have normal distribution of our data.

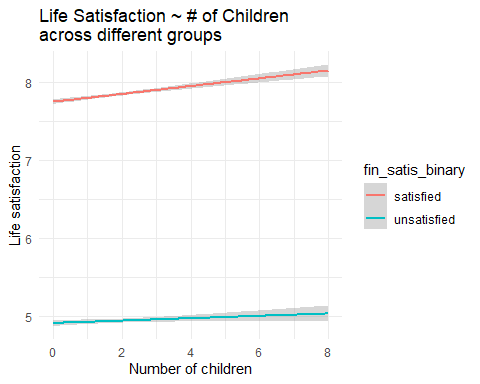
# Problem 3. Linear Models

## Divide dataset by fsatis and plot the relationship

divided\_data <- wvs\_data3 %>%   
 drop\_na() %>%   
 mutate(fin\_satis\_binary = ifelse(fsatis >= 6, "satisfied", "unsatisfied"))  
  
head(divided\_data, 5)

## # A tibble: 5 x 10  
## index wave country satis chldr fsatis sex age emp fin\_satis\_binary  
## <chr> <chr> <chr> <dbl> <dbl> <dbl> <fct> <dbl> <fct> <chr>   
## 1 38144 3 Albania 3 0 6 0 18 0 satisfied   
## 2 38145 3 Albania 9 0 7 1 18 0 satisfied   
## 3 38147 3 Albania 7 0 5 1 26 1 unsatisfied   
## 4 38149 3 Albania 7 2 3 0 38 1 unsatisfied   
## 5 38150 3 Albania 2 0 3 1 35 1 unsatisfied

ggplot(divided\_data, aes(chldr, satis, color = fin\_satis\_binary)) +  
 geom\_smooth(method = "lm") +  
 xlab("Number of children") +  
 ylab("Life satisfaction") +  
 ggtitle("Life Satisfaction ~ # of Children\nacross different groups") +  
 theme\_minimal()



On the basis of the visualization, we can observe that among those who are financially satisfied the initial level of life satisfaction is much higher than among those who are not financially satisfied. Moreover, we see a more pronounced positive trend with of life satisfaction the increase in the number of children among those who are *financially satisfied* than among those who are not.

## Run a regression model (or a set of regression models) that tests whether the relationship between life satisfaction and the number of children is different

#turn off scientific notation  
options(scipen = 999)  
  
for\_lm <- divided\_data %>%   
 split(.$fin\_satis\_binary)  
  
class(for\_lm) # list with two dataframes for two groups

## [1] "list"

fit\_satis <- lm(satis ~ chldr + age + sex + emp, data = for\_lm[["satisfied"]])  
fit\_unsatis <- lm(satis ~ chldr + age + sex + emp, data = for\_lm[["unsatisfied"]])  
  
summary(fit\_satis); summary(fit\_unsatis)

##   
## Call:  
## lm(formula = satis ~ chldr + age + sex + emp, data = for\_lm[["satisfied"]])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.1654 -0.8714 0.1797 1.2397 2.3060   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.7736618 0.0335092 231.986 < 0.0000000000000002 \*\*\*  
## chldr 0.0522386 0.0069134 7.556 0.0000000000000427 \*\*\*  
## age -0.0005541 0.0007429 -0.746 0.456   
## sex1 -0.0286628 0.0215641 -1.329 0.184   
## emp1 0.0104450 0.0221952 0.471 0.638   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.797 on 29698 degrees of freedom  
## Multiple R-squared: 0.002331, Adjusted R-squared: 0.002197   
## F-statistic: 17.35 on 4 and 29698 DF, p-value: 0.00000000000003161

##   
## Call:  
## lm(formula = satis ~ chldr + age + sex + emp, data = for\_lm[["unsatisfied"]])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.8053 -1.8061 -0.0974 1.7434 6.1121   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.622139 0.043794 128.376 < 0.0000000000000002 \*\*\*  
## chldr 0.097429 0.008703 11.195 < 0.0000000000000002 \*\*\*  
## age -0.021502 0.000915 -23.500 < 0.0000000000000002 \*\*\*  
## sex1 -0.035603 0.026349 -1.351 0.177   
## emp1 0.107199 0.027071 3.960 0.0000751 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.395 on 34637 degrees of freedom  
## Multiple R-squared: 0.0183, Adjusted R-squared: 0.01819   
## F-statistic: 161.4 on 4 and 34637 DF, p-value: < 0.00000000000000022

**NB! The remark p-value < 0.05 in the paranthesis means that the predictor is statistically significant**

For the sub-sample of those who are financially satisfied, we observe that the average level of life satisfaction is 7.77 if there is no children in the family and all else equal. An increase in the number of children by 1 child, the average level of life satisfaction increases by 0.05 units (p-value < 0.05). An increase in the age will decrease the level of life satisfaction by 0.0005 on average (p-value < 0.05). Being a male (and all else equal) will state that the average level of life satisfaction is 7.74 (Intercept - sex1); and being a female will state that the average level of life satisfaction is 7.77. The variable for sex is *statistically insignificant*. Employment is statistically significant in terms of level of life satisfaction: having a job guarantee that a respondent has 7.78 units of life satisfaction on average. However, the R-squared (and adjusted R-squared) are very low - 0.2%.

For the sub-sample of those who are *not* financially satisfied, we see that the average level of life satisfaction equals to 5.62 if there is no children in a family and all else equal. In addition, the increase of the number of children by 1 child will increase the average level of life satisfaction by 0.09 units (p-value < 0.05). The increase in age by 1 year will decrease the average level of life satisfaction by 0.02 units (p-value < 0.05). Being a male (and all else equal) guarantee the average level of 5.58 units; females are much satisfied with their lives on average - 5.62. However, sex is *statistically insignificant* predictor. Having a job will statistically significant guarantee the average level of life satisfaction of 5.72 (p-value < 0.05). Adjusted R-squared is low - 1.8% - but, nevertheless, higher than in the other sub-sample model.

# Problem 4. Run a regression model that tests the moderation effect of fsatis on the relationship between life satifaction and the number of children

fit\_interaction <- lm(satis ~ chldr + fin\_satis\_binary + chldr:fin\_satis\_binary + age + sex + emp, data = divided\_data)  
  
summary(fit\_interaction)

##   
## Call:  
## lm(formula = satis ~ chldr + fin\_satis\_binary + chldr:fin\_satis\_binary +   
## age + sex + emp, data = divided\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.3498 -1.3926 0.0654 1.4137 5.6672   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 8.1067297 0.0306302 264.664  
## chldr 0.1027406 0.0077513 13.255  
## fin\_satis\_binaryunsatisfied -2.8001233 0.0254717 -109.931  
## age -0.0119537 0.0006016 -19.870  
## sex1 -0.0294714 0.0173913 -1.695  
## emp1 0.0722056 0.0178771 4.039  
## chldr:fin\_satis\_binaryunsatisfied -0.0417497 0.0102458 -4.075  
## Pr(>|t|)   
## (Intercept) < 0.0000000000000002 \*\*\*  
## chldr < 0.0000000000000002 \*\*\*  
## fin\_satis\_binaryunsatisfied < 0.0000000000000002 \*\*\*  
## age < 0.0000000000000002 \*\*\*  
## sex1 0.0902 .   
## emp1 0.0000537 \*\*\*  
## chldr:fin\_satis\_binaryunsatisfied 0.0000461 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.145 on 64338 degrees of freedom  
## Multiple R-squared: 0.3151, Adjusted R-squared: 0.315   
## F-statistic: 4933 on 6 and 64338 DF, p-value: < 0.00000000000000022

The average level of life satisfaction when there are not children and all else equal is 8.10. The increase in children by 1 child will increase the average level of life satisfaction by 0.10 (p-value < 0.05) when a person is financially satisfied and all else equal. The increase in age by 1 year will decrease the average level of life satisfaction by 0.011 units (p-value < 0.05). Sex is statistically insignificant in the model. Being unsatisfied with the financial conditions (and all else equal) predicts the average level of life satisfaction on 5.3 (p-value < 0.05).

Being employed (and all else equal) increases the average level of life satisfaction by 8.17 units. The interaction with financial satisfaction (p-value < 0.05) can be interpreted as follows:

d(Life Satisfaction) / d(Number of children) = (beta\_1 \* Number of Children)' + (beta\_6 \* Interaction)'

And we get:

d(Life Satisfaction) / d(Number of children) = 0.102740 + (-0.0417497) = 0.0609903

This value means that the increase in number of children when you are unsatisfied with your financial conditions (and all else equal) will increase the average level of life satisfaction by 0.06 units.

# Problem 5. Explore how the marginal effect of the number of children and its significance depends on the values of satisfaction with the household financial situation

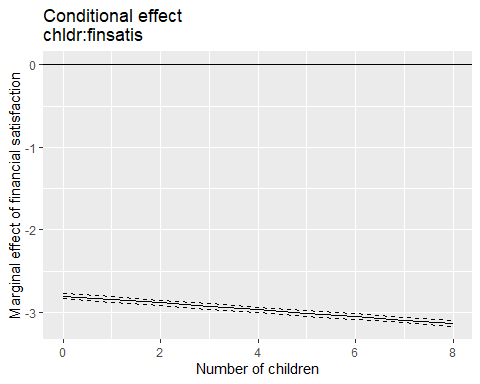
me <- margins(fit\_interaction)  
  
summary(me)

## factor AME SE z p lower upper  
## age -0.0120 0.0006 -19.8701 0.0000 -0.0131 -0.0108  
## chldr 0.0803 0.0057 14.1686 0.0000 0.0692 0.0914  
## emp1 0.0722 0.0179 4.0390 0.0001 0.0372 0.1072  
## fin\_satis\_binaryunsatisfied -2.8774 0.0170 -169.2710 0.0000 -2.9108 -2.8441  
## sex1 -0.0295 0.0174 -1.6946 0.0902 -0.0636 0.0046

The average marginal effect (further - AME) for age equals -0.12; for our main predictor - 0.08; for employment variable - 0.072; for being in the group of financially unsatisfied - -2.8774. These mentioned predictors are statistically significant (p-value < 0.05).

The AME for sex (male) is -0.0295 but it is statistically insignificant (p-value > 0.05).

ggplot(me, aes(x = chldr)) +   
 geom\_line(aes(y = dydx\_fin\_satis\_binaryunsatisfied)) +  
 geom\_line(aes(y = dydx\_fin\_satis\_binaryunsatisfied + 1.96 \* sqrt(Var\_dydx\_fin\_satis\_binaryunsatisfied)), linetype = 2) +  
 geom\_line(aes(y = dydx\_fin\_satis\_binaryunsatisfied - 1.96 \* sqrt(Var\_dydx\_fin\_satis\_binaryunsatisfied)), linetype = 2) +  
 geom\_hline(yintercept = 0) +  
 ggtitle("Conditional effect\nchldr:finsatis") +  
 xlab("Number of children") + ylab("Marginal effect of financial satisfaction")



We observe that our line does not cross the y = 0. Also, we see that confidence interval for our AME is very narrow. It gives an insight that our predictor of number of children affects the life satisfaction when interacted with financial satisfaction as a moderator.